

DOI: 10.20103/j.stxb.202408211971

李春辉, 欧阳逸云, 何燕, 倪荣雨, 曾爱聪, 苏漳文, 郭福涛. 基于空间广义加性模型的黑龙江省林火发生预测. 生态学报, 2025, 45(8): 3957-3968.  
Li C H, Ouyang Y Y, He Y, Ni R Y, Zeng A C, Su Z W, Guo F T. Forest fire occurrence prediction in Heilongjiang Province based on spatial generalized additive models. Acta Ecologica Sinica, 2025, 45(8): 3957-3968.

## 基于空间广义加性模型的黑龙江省林火发生预测

李春辉<sup>1,2</sup>, 欧阳逸云<sup>1,2</sup>, 何燕<sup>1,2</sup>, 倪荣雨<sup>1,2</sup>, 曾爱聪<sup>1,2</sup>, 苏漳文<sup>1,3</sup>, 郭福涛<sup>1,2,\*</sup>

1 福建农林大学, 福州 350002

2 福建农林大学 3S 技术与资源优化利用福建省高校重点实验室, 福州 350002

3 漳州职业技术学院, 漳州 363000

**摘要:** 林火对森林生态系统有着重大影响, 造成了广泛的生态破坏和重大的经济损失, 因此建立准确可靠的预测模型对森林火灾防控至关重要。研究旨在对比分析 Logistic 回归模型和空间广义加性模型在林火发生预测和火险等级划分方面的应用效果, 为森林火灾防控提供更科学的模型依据。选取 2006—2020 年的黑龙江省林火数据, 结合气象、地形、植被等多种影响因素, 对 Logistic 回归模型和四种不同基函数的空间广义加性模型进行评估。结果显示: 相较于传统 Logistic 回归模型, 由高斯过程平滑样条基 (GP), 三次样条基 (CR), 薄板回归样条基 (TP), 自适应样条基 (AD) 拟合的空间广义加性模型均展现出更优异的拟合效果和预测能力。其中, AD 拟合的空间广义加性模型效果最佳, 其测试集准确率提高 4.2%, AUC 值提升 0.053。模型预测显示, 黑龙江省的高火险区主要分布在西北和中南地区, 与该省实际的防火布局高度吻合。研究表明, 空间信息在森林火灾发生预测中具有显著作用。同时, 基于自适应样条基的空间广义加性模型能够对自变量进行分段线性解释, 为黑龙江省制定精准的火灾预防措施、优化消防资源配置提供了更具针对性的理论参考和决策支持。

**关键词:** 林火预测模型; Logistic 回归模型; 空间广义加性模型; 分段效应; 平滑样条函数

## Forest fire occurrence prediction in Heilongjiang Province based on spatial generalized additive models

LI Chunhui<sup>1,2</sup>, OUYANG Yiyun<sup>1,2</sup>, HE Yan<sup>1,2</sup>, NI Rongyu<sup>1,2</sup>, ZENG Aicong<sup>1,2</sup>, SU Zhangwen<sup>1,3</sup>, GUO Futao<sup>1,2,\*</sup>

1 Fujian Agriculture and Forestry University, Fuzhou 350002, China

2 3S Technology and Resource Optimization Utilization Key Laboratory of Fujian Universities, Fujian Agriculture and Forestry University, Fuzhou 350002, China

3 Zhangzhou Institute of Technology, Zhangzhou 363000, China

**Abstract:** Forest fires significantly impact on forest ecosystems, causing extensive ecological damage and significant economic losses. This underscores the urgent need for accurate and reliable forecasting models to support effective forest fire prevention and management strategies. This study compares the performance of Logistic Regression and Spatial Generalized Additive Models in predicting forest fire occurrences and accessing fire risk levels, thus providing a solid scientific basis for informed decision-making in fire prevention efforts. Using comprehensive forest fire data from Heilongjiang Province, collected over a 15-year period (2006—2020), the study integrated multiple influencing factors, including meteorological conditions, topography, and vegetation characteristics, to assess the predictive capabilities of the Logistic Regression Model and four different basis functions of the Spatial Generalized Additive Models. The SGAMs tested in this study included models with Gaussian process smooth spline (GP), Cubic regression spline (CR), Thin plate regression spline (TP), and Adaptive smooth spline (AD). The results clearly demonstrated the significant advantages of Spatial Generalized Additive

**基金项目:** 国家重点研发计划战略性国际科技创新合作重点专项 (2018YFE0207800)

**收稿日期:** 2024-08-21; **网络出版日期:** 2025-01-20

\* 通讯作者 Corresponding author. E-mail: guofutao@126.com

Models over the conventional Logistic Regression Model, with Spatial Generalized Additive Models consistently achieving superior model fit and predictive accuracy. Among these models, the Adaptive smooth spline (AD) model exhibited the best overall performance, with a 4.2% increase in test set accuracy and a 0.053 improvement in the area under the curve (AUC) compared to its counterparts. Further spatial analysis identified high-risk forest fire areas in Heilongjiang Province, primarily concentrated in the northwest and central-southern regions. These findings closely align with the province's existing fire prevention layout, underscoring the practical relevance of Spatial Generalized Additive Models-based models for real-world applications. Notably, the study highlights the critical role of spatial data in forest fire forecasting. By incorporating spatial data, the understanding of fire risk patterns across diverse landscapes is significantly enhanced, enabling region-specific prevention strategies and optimized resource allocation. Additionally, the unique advantages of the AD model stand out. Its ability to provide segmented linear insights into the effects of influencing variables allows for a more nuanced interpretation of how factors such as meteorology, topography, and vegetation interact to influence fire risk. This capability supports the development of more precise and targeted fire prevention measures. In conclusion, this research underscores the potential of Spatial Generalized Additive Models, particularly the AD model, as highly effective tools for advancing forest fire management. By adopting these advanced models, regions like Heilongjiang Province can strengthen fire prevention strategies, better protect forest ecosystems, and achieve sustainable forest management goals while minimizing the risks of future catastrophic disasters.

**Key Words:** forest fire prediction model; logistic regression model; spatial generalized additive model; segmental effect; smoothing spline function

林火是森林生态系统的重要干扰因素,也是一种严重的自然灾害,给人类社会和生态环境带来了巨大的损失<sup>[1-2]</sup>。因此,准确预测林火的发生对于采取及时有效的控制措施至关重要。然而,鉴于林火的复杂性和不确定性,现有的林火发生预测模型在准确性和适应性方面仍面临诸多挑战。

目前,国内外针对林火发生预测模型进行了大量研究,主要方法包括泊松回归、负二项回归和 Logistic 回归模型等参数回归模型<sup>[3-6]</sup>,以及人工神经网络、贝叶斯网络、随机森林算法等机器学习算法<sup>[7-10]</sup>。其中, Logistic 回归模型在林火预测中被广泛应用,并取得了一定的效果<sup>[6,11-12]</sup>。然而,森林火灾是一个复杂的非线性过程,受到多种因素的影响,这些因素之间的关系错综复杂<sup>[13-14]</sup>。传统的 Logistic 回归模型无法直接处理变量之间的非线性关系,只能通过引入交互项或多项式项来近似非线性关系,但这种方法可能会增加模型的复杂度<sup>[15]</sup>。在机器学习算法中,结构简单的模型通常具备良好的可解释性,但拟合能力较弱,准确性不高<sup>[16]</sup>;相反,结构复杂的模型虽然能够有效处理大量林火数据和非线性关系,但参数众多、机制复杂且缺乏透明性,因此可解释性较差<sup>[17]</sup>。

本研究所采用的空间广义加性模型 (Spatial generalized additive model, SGAM) 是基于广义加性模型 (Generalized additive model, GAM) 的扩展,在建模时考虑了空间相关性。相较于 Logistic 回归模型, SGAM 在处理具有空间结构的数据时更为适用<sup>[18]</sup>,且通过链接函数来建立解释变量和响应变量之间的复杂关系,更适用于拟合高度非线性的数据<sup>[19]</sup>。SGAM 可以通过可视化平滑函数的形状来理解每个预测变量的分段效应和强度<sup>[20]</sup>,这有助于深入理解林火的空间分布规律,从而为林火风险评估和精细化管理提供科学依据。目前, GAM 已广泛应用于全球森林火灾预测研究。最早由 Brillinger 等<sup>[21]</sup>用于预测俄勒冈州的火灾数量并取得了良好的预测效果; Parisien 等<sup>[22]</sup>证明 GAM 能够有效预测美国西部地区林火发生的空间异质性; Woolford 等<sup>[23]</sup>则在加拿大安大略省通过响应依赖采样和建模技术,开发了精细时空尺度的林火预测模型; Phelps 等<sup>[24]</sup>通过对比较准后的统计模型和机器学习模型,发现 GAM 与机器学习模型具有相似的性能,且 GAM 在可解释性上更具优势。以往研究主要集中在—维参数分析上,而本文结合了二维空间参数和—维参数展开研究,进一步探讨了数据之间的复杂关系。

本文以林火频发的黑龙江省为研究对象<sup>[13,25]</sup>,在文献综述的基础上综合考虑了常见的林火驱动因素,并

将空间位置信息纳入预测变量。研究中分别使用 Logistic 回归模型和空间广义加性模型建立黑龙江省林火发生预测模型。根据预测结果和模型阈值分类标准,划分黑龙江省的火险等级。通过对模型的拟合与检验,评估了空间广义加性模型在林火预测中的可行性,为林火预测模型的发展和应用提供新的思路。

## 1 材料与方法

### 1.1 研究区概况

黑龙江省地处我国最东北部(43°26′—53°33′N, 121°11′—135°05′E),属寒温带大陆性季风气候,年降水量 400—800mm,年日照时数为 2400—2800h。全省陆地面积  $47.3 \times 10^4 \text{ km}^2$ ,森林面积达  $1990.46 \times 10^4 \text{ hm}^2$ ,森林蓄积量  $18.47 \times 10^8 \text{ m}^3$ ,森林覆盖率达 43.8%,是我国森林火灾高发地区之一。主要分布有红松(*Pinus koraiensis*)、樟子松(*Pinus sylvestris var. mongolica*)和白桦(*Betula platyphylla*)等易燃树种以及青楷槭(*Acer tegmentosum*)等防火树种。根据中国统计年鉴(<http://www.stats.gov.cn/sj/ndsjsj>),2006—2020 年间,黑龙江省共发生火灾 818 起,其中重大火灾 8 起,火场总面积  $595253 \text{ hm}^2$ ,其中的受灾森林面积为  $325973 \text{ hm}^2$ ,火灾共造成 43 人伤亡。

### 1.2 数据来源

#### 1.2.1 火点数据

本研究采用 2006—2020 年期间的林火数据,来源于美国国家航天航空局(NASA)提供的每日林火产品 MCD14DL,空间分辨率为 1km,数据包含火灾位置、探测日期和时间、置信度、火灾辐射率等信息。MCD14DL 是基于 MODIS 传感器的数据产品,MODIS 搭载于 Terra 和 Aqua 卫星,近年来广泛应用于林火研究<sup>[26]</sup>。本研究选取置信度大于 70%的火点。原因在于,一方面,卫星过境时间与实际火灾发生时间之间可能存在差异,导致一些短时燃烧的火灾未能被探测到<sup>[27]</sup>;另一方面,在高植被覆盖区、浓密云层或烟雾下,火灾产生的热量可能无法被卫星传感器有效识别,导致置信度被低估<sup>[28]</sup>,因此不应选取置信度过高的火点。结合中国 1:10 万比例尺土地利用现状遥感监测数据库提供的 1km 分辨率土地利用类型图,剔除非林地火点<sup>[25]</sup>,最终得到可信火点 10479 个,如图 1。由于 Logistic 回归模型和 SGAM 遵循二项分布,本研究利用 ArcGIS 软件按照 1:1 的比例生成随机点作为非火点,为避免与已知火点重叠,设置任意两个随机点的最小距离为 1km,且随机点在空间和时间上的分布完全随机<sup>[6]</sup>。

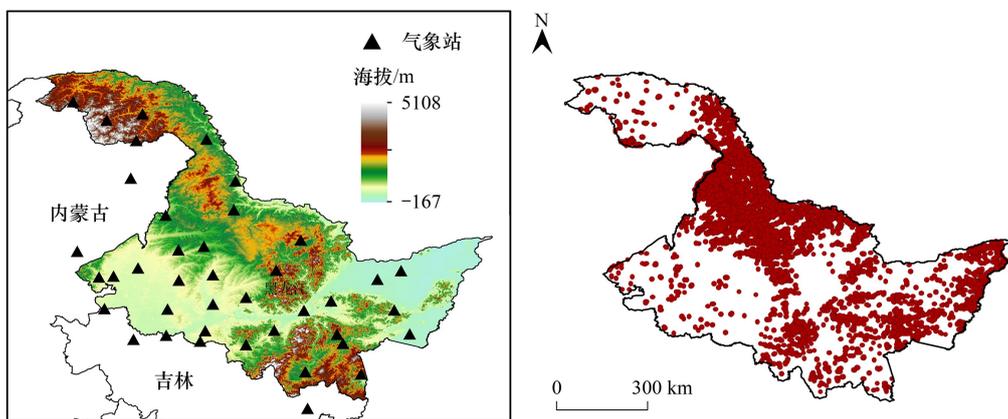


图 1 研究地区火点和气象站点的分布

Fig.1 Distribution of fire points and meteor stations in the study area

#### 1.2.2 气象数据

本研究所用的日气象数据来源于中国气象数据网(<http://data.cma.cn>)提供的 2006—2020 年间黑龙江省内及周边省份 39 个国家级气象站的观测记录。从中选取当日最高气温、日照时数、相对湿度、降水等气象因

素参与模型的拟合。这些气象指标的变化会影响森林燃料的含水率和着火条件,从而影响森林火灾的发生<sup>[29]</sup>。使用 ArcGIS 的连接功能,提取距离火点和随机点最近的气象站数据,并通过区站号作为唯一字段提取对应的日气象数据。

### 1.2.3 地形数据

地形因素会影响燃料的分布并干扰气象条件<sup>[30]</sup>。本研究的海拔数据来源于地理空间数据云(<http://www.gscloud.cn>)提供的分辨率为 25m 的数字高程模型(DEM),利用 ArcGIS 的 3D 分析功能计算坡度,提取火点和随机点的海拔与坡度信息。

### 1.2.4 植被覆盖

归一化植被指数(Normalized difference vegetation index, NDVI)作为衡量燃料状况的重要指标,是影响火灾发生的关键因素之一。逐月 NDVI 栅格数据来源于国家青藏高原科学数据中心(<http://data.tpdc.ac.cn>),使用的是 2006—2020 年间卫星传感器 MOD13A3 提供的 250m 分辨率逐月 NDVI 数据。

### 1.2.5 数据统计

本研究共包含 17674 个样本点,并将数据集按 7:3 比例随机划分为训练集和测试集。其中,训练集包含 12370 个样本,测试集包含 5304 个样本。表 1 展示了两个数据集中各输入变量的参数统计。

表 1 自变量统计  
Table 1 Independent variables statistics

变量名称 Variable name	建模数据 Fitting data					检验数据 Validation data				
	平均值 Mean	最小值 Minimum	最大值 Maximum	标准差 Standard deviation	变异系数 Coefficient of variation/%	平均值 Mean	最小值 Minimum	最大值 Maximum	标准差 Standard deviation	变异系数 Coefficient of variation/%
纬度 Latitude/(°)	48.91	43.70	53.39	1.82	3.7	48.88	43.70	53.39	1.80	3.7
经度 Longitude/(°)	127.92	121.70	134.75	2.66	2.1	127.94	121.89	134.75	2.67	2.1
最高温度 Maximum temperature/°C	18.39	0.20	43.70	8.55	46.5	18.22	1.30	41.60	2.67	2.1
日照时数 Sunshine duration/h	8.09	0.00	15.30	3.47	42.9	8.10	0.00	15.70	3.47	42.8
相对湿度 Relative humidity/%	56.31	12.00	100	17.45	31.0	56.35	12.00	100.00	17.54	31.1
降水 Precipitation/mm	0.76	0.00	92.30	3.85	507.2	0.74	0.00	84.40	3.79	512.6
海拔 Elevation/m	345.56	15.00	1299.00	176.50	51.1	347.72	9.00	1299.00	179.02	51.5
坡度 Slope/(°)	4.25	0.00	32.64	3.61	85.1	4.33	0.00	31.34	3.68	84.8
植被归一化指数 Normalized difference vegetation index	0.44	0.0003	0.98	0.23	52.6	0.44	0.0013	0.97	0.23	51.5

## 1.3 研究方法

### 1.3.1 变量筛选

首先,使用方差膨胀因子(Variance inflation factor, VIF)检验解释变量之间的多重共线性,并排除具有多重共线性的因子(VIF>10)。然后,通过双向逐步回归方法去除影响不显著的因子,最终将剩余的解解释变量纳入模型的拟合。

### 1.3.2 Logistic 回归模型

Logistic 回归模型是一种用于二分类问题的回归模型,广泛应用于林火发生的预测与分析领域。假设二分类目标变量  $Y$  表示林火发生( $Y=1$ )与不发生( $Y=0$ ),林火发生的概率为  $P$ ,不发生的概率为  $1-P$ ,则  $P$  与输入变量  $X_i(i=1,2,\dots,n)$  之间的回归关系为<sup>[11]</sup>:

$$\text{logit}(P) = \ln \frac{P}{1-P} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

经过变换得到的 Logistic 回归模型为:

$$P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)]}$$

式中,  $P$  为林火发生的概率,  $n$  为输入变量的总数,  $\beta_0$  是模型的截距项,  $\beta_1, \beta_n$  是模型的回归系数,  $X_1, X_n$  是输入变量的值。

### 1.3.3 空间广义加性模型

空间广义加性模型(Spatial Generalized Additive Model, SGAM)是在广义加性模型的基础上,结合空间统计的扩展模型,通过将自变量的影响建模为非线性、非参数化的平滑函数,捕捉自变量与响应变量之间的复杂关系<sup>[18]</sup>。

在空间数据中,变量( $X, Y$ )的空间位置由  $s = (s_1, s_2)$  表示,  $s_1$  表示经度,  $s_2$  表示纬度。由于因变量遵循二项分布,本研究使用 Logistic SGAM 模型,其二元响应式为<sup>[18]</sup>:  $P(Y=1|s, X)$

$$= \frac{\exp(\alpha + g(s_1, s_2) + f_1(X_1) \dots + f_p(X_p))}{1 + \exp(\alpha + g(s_1, s_2) + f_1(X_1) \dots + f_p(X_p))}$$

式中,  $\alpha$  是常数,  $g()$  是双变量空间效应的平滑函数,并且  $f_j()$  ( $j=1, \dots, P$ ) 是平滑样条,其说明了火灾发生的概率和解释变量  $X_j$  之间的非线性关系。它与 GAM 模型的不同之处在于,它通过函数  $g = (s_1, s_2)$  将平面坐标作为模型输入,双变量项  $g = (s_1, s_2)$  使用 2D 样条来近似。

每个平滑函数  $f_j()$  的表达式为<sup>[31]</sup>:

$$f_j(X_j) = \sum_{k=1}^K \beta_{j,k} b_{j,k}(X_j)$$

式中,  $k$  为基函数(一组可以用来表示复杂非线性函数的简单函数)的数量,  $b_{j,k}$  为  $k$  个  $X_j$  的基函数,  $\beta_{j,k}$  为相应的系数。

SGAM 由参数项和包含二维平滑器及一维平滑样条的非参数项构成<sup>[18]</sup>。本研究使用二维薄板样条平滑来拟合空间交互作用,这是因为经度和纬度的单位相同,惩罚矩阵的计算和光滑度  $\lambda$  应用相同的标准<sup>[32]</sup>。针对一维平滑项,本文参考 Simpson 的研究<sup>[33]</sup> 选择了两种常见的样条基:三次回归样条(Cubic regression spline, CR)和薄板回归样条(Thin plate regression spline, TP)以及两种特殊类型的样条基:高斯过程平滑样条(Gaussian process smooth spline, GP)和自适应样条(Adaptive smooth spline, AD)对 SGAM 的一维平滑项进行拟合。

模型系数和平滑参数通过最大似然估计(Maximum Likelihood, ML)进行估计。尽管加性模型通常使用限制性最大似然估计(Restricted Maximum Likelihood, REML),但当模型包含多个解释变量,应采用 ML 估计。由于本研究涉及多维解释变量,因此选用了 ML 方法来估计模型系数和平滑参数<sup>[33]</sup>。

### 1.3.4 模型评价和检验

本研究通过 5 种指标对模型进行评价:受试者工作特征(Receiver operating characteristic, ROC)曲线下的面积(Area under the curve, AUC)、平均绝对误差(Mean absolute error, MAE)、均方误差(Mean square error, MSE)和均方根误差(Root mean squared error, RMSE)。其中, AUC 的取值范围通常是 0.5—1,越接近 1 表示模型效果越好<sup>[12]</sup>。MAE 表示预测值与实际值之间绝对误差的平均值,反映了残差的平均水平。MSE 是预测值与实际值差值平方的平均值,衡量残差的方差。RMSE 是 MSE 的平方根,比 MSE 对大误差更加敏感,反映残差的标准偏差。相较于 MAE, MSE 和 RMSE 更注重惩罚大的预测误差。上述指标越低,模型的预测精度越高。公式如下:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_0|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_0)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_0)^2}$$

式中,  $n$  为样本个数,  $y_i$  为实际值,  $y_0$  为预测值。

### 1.3.5 模型预测精度评价

本研究使用 ROC 曲线中的敏感度和特异度计算约登指数 (Youden index) 作为预测林火发生与否的最佳切割值 (Cut-off point)。约登指数为敏感度与特异度之和减去 1, 取值范围为 0—1, 代表分类模型预测林火发生与不发生的总能力。根据约登指数的值可以确定判断模型预测精度的最佳阈值, 大于该阈值的模型预测概率表示林火发生; 小于该阈值的模型预测概率表示林火不发生。

模型的预测性能还可以通过准确率 (Accuracy) 来衡量:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

式中,  $TP$  表示真阳性, 火点预测为真的数据;  $TN$  表示真阴性, 非火点 (随机点) 预测为真的数据;  $FP$  表示假阳性, 火点预测为假的数据;  $FN$  表示假阴性, 非火点 (随机点) 预测为假的数据。

在本研究中利用 R 语言来拟合和验证模型, Logistic 回归模型的拟合由 `stats` 包的 `glm` 函数得到, SGAM 的拟合由 `mgcv` 包的 `gam` 函数得到。

## 2 结果与分析

### 2.1 Logistic 回归模型的构建

本研究所有变量均通过方差膨胀因子 (VIF) 的多重共线性检验 ( $VIF < 5$ ), 使用逐步回归方法排除了一个不显著的变量 (坡度), 其余变量进入基础模型的拟合。该模型的回归参数见表 2。

表 2 Logistic 回归模型的参数估计

Table 2 Parameter estimation of Logistic regression model

输入变量 Input variable	估计值 Estimate value	标准误差 Standard error	$P$	输入变量 Input variable	估计值 Estimate value	标准误差 Standard error	$P$
Intercept	0.350	0.028	<0.001	Preci	-1.059	0.011	<0.001
Max_temp	-0.735	0.026	<0.001	Elev	0.136	0.022	<0.001
SSD	0.329	0.028	<0.001	NDVI	0.414	0.027	<0.001
RHU	-1.137	0.031	<0.001				

Max\_temp: 最高温度 Maximum temperature; SSD: 日照时数 Sunshine duration; RHU: 相对湿度 Relative humidity; Preci: 降水 Precipitation; Elev: 海拔 Elevation; NDVI: 归一化植被指数 Normalized difference vegetation index

### 2.2 空间广义加性模型的构建

使用四种平滑样条基拟合 SGAM 的结果如表 3 所示, 模型中的 7 个非参数项在各样条基拟合中均表现出显著性 ( $P$ -value < 0.1)。

通过对空间信息的交互作用进行可视化 (图 2), 发现空间坐标的交互作用对研究区域森林火灾的发生具有显著且波动的影响, 林火高发区位于研究区的左下角。

图 3 展示了使用四种平滑样条基拟合的 SGAMs 中各解释变量的部分依赖性图。置信区间的宽度随着解释变量观测值的减少而增大。虽然各平滑样条基拟合的部分依赖图在趋势上存在细微差异, 但所有模型中解释变量的趋势表现出高度一致性。同时各因子对森林火灾发生的分段效应进一步验证了非参数模型在揭示自变量与森林火灾发生之间关系方面的适用性。其中, 降水与林火发生呈显著的全局负相关。相对湿度和海拔与林火发生的关系相似: 两者在低值时与林火发生呈显著正相关, 当相对湿度超过约 55%, 海拔超过约 300m 时, 转为显著负相关。NDVI 与林火发生的关系呈现非线性; 当 NDVI 小于约 0.22 时与林火发生呈显著负相关, 在 0.22—0.85 之间呈显著正相关, 超过 0.85 后, 又转为负相关。森林火灾的发生与最高温度、日照时

数之间的关系比较复杂。

表 3 基于 GP,CR,TP,AD 拟合的 SGAMs 的回归参数

Table3 Regression parameters of SGAMs based on GP,CR,TP,AD

模型 Model	参数项 Parametric item				非参数项 Nonparametric item				
	截距 Intercept	估计值 Estimated	标准误差 Std.error	<i>P</i>	平滑项 Smoothing term	<i>df</i>	卡方检验 Chi-square	<i>P</i>	
平滑样条 Smoothing spline	$\beta_1$	-0.135	0.095	0.154	$S_1$ (Latitude,Longitude)	15.689	63.902	<0.001	
高斯过程平滑样条 Gaussian process smooth					$S_1$ (Max_temp)	6.850	360.048	<0.001	
spline(GP)					$S_1$ (SSD)	7.189	250.617	<0.001	
					$S_1$ (RHU)	1.000	600.777	<0.001	
					$S_1$ (Preci)	4.703	135.237	<0.001	
					$S_1$ (Elev)	1.000	8.066	<0.0	
三次回归样条 Cubic regression	$\beta_2$	-0.153	0.100	0.128	$S_2$ (Latitude,Longitude)	14.868	52.75	<0.001	
					spline(CR)	$S_2$ (Max_temp)	6.856	348.01	<0.001
					$S_2$ (SSD)	7.510	247.48	<0.001	
					$S_2$ (RHU)	3.281	593.60	<0.001	
					$S_2$ (Preci)	5.127	146.63	<0.001	
					$S_2$ (Elev)	3.437	16.34	<0.001	
薄板回归样条 Thin plate regression	$\beta_3$	-0.159	0.101	0.115	$S_3$ (Latitude,Longitude)	14.900	51.74	<0.001	
					spline(TP)	$S_3$ (Max_temp)	6.907	356.83	<0.001
					$S_3$ (SSD)	7.147	250.57	<0.001	
					$S_3$ (RHU)	1.001	602.79	<0.001	
					$S_3$ (Preci)	4.799	136.03	<0.001	
					$S_3$ (Elev)	2.340	11.75	<0.01	
自适应样条 Adaptive smoothing	$\beta_4$	-0.126	0.109	0.249	$S_4$ (Latitude,Longitude)	15.425	59.74	<0.001	
					spline(AD)	$S_4$ (Max_temp)	11.053	356.61	<0.001
					$S_4$ (SSD)	15.772	273.49	<0.001	
					$S_4$ (RHU)	2.943	540.14	<0.001	
					$S_4$ (Preci)	5.016	195.65	<0.1	
					$S_4$ (Elev)	4.187	17.43	<0.001	
					$S_4$ (NDVI)	12.748	490.86	<0.001	

$\beta_1$ — $\beta_4$ 分别表示 GP,CR,TP,AD 对应的截距, $S_1$ — $S_4$ 分别表示 GP,CR,TP,AD 对应的平滑样条基

### 2.3 模型检验

表 4 展示了不同模型检验结果的对比。可以看出,四种平滑样条基拟合的 SGAMs 的 AUC 值和准确率均高于 Logistic 回归模型。在这四种 SGAMs 中,由 AD 平滑样条基拟合的 SGAM 预测准确率明显优于其它三种 SGAMs。其训练集准确率比 Logistic 回归模型提高了 4.2%,测试集准确率提高了 4.1%,具有更优的拟合精度。为了进一步评估 SGAM 和 Logistic 回归模型在林火预测中的精度,图 4 给出了 Logistic 回归模型和 AD-SGAM 的残差分布图,与 Logistic 回归模型相比,AD-SGAM 的残差分布更集中在 0 附近,说明其预测结果更加准确。表 5 是两种模型各项检验指标的对比结果,AD-SGAM 的 MAE、MSE 以及 RMSE 值均低于 Logistic 回归模型,进一步证明由 AD 拟合的 SGAM 的预测精度高于 Logistic 回归模型。

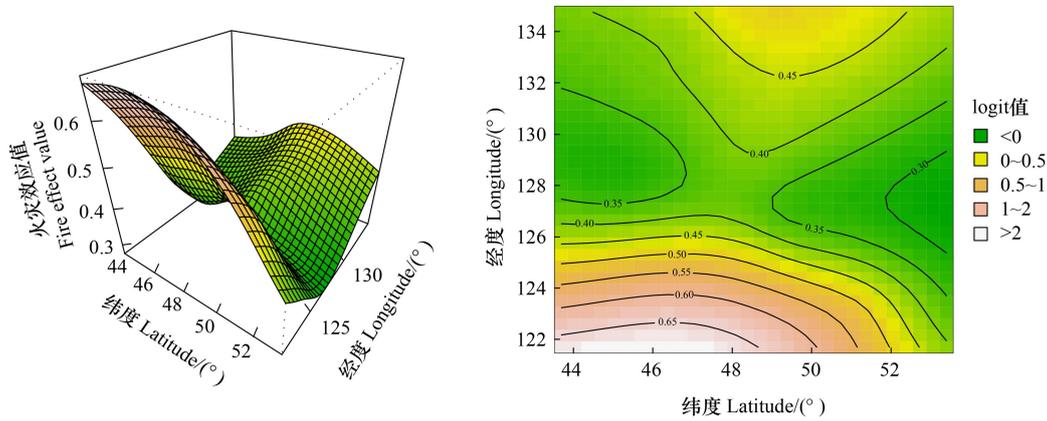


图 2 空间交互作用对火灾效应的影响

Fig.2 The effect of spatial interaction on fire effect

logit:对数几率 Logarithm of odds

表 4 模型检验结果的对比

Table 4 Comparison of model validation results

模型 Model	阈值 Threshold value	曲线下面积 Area under the curve	准确率 Accuracy/%	
			训练集 Training set	测试集 Test set
Logistic 回归模型	0.475	0.816	78.4	77.4
GP-SGAM	0.628	0.864	82.5	81.5
CR-SGAM	0.622	0.866	82.5	81.3
TP-SGAM	0.630	0.864	82.5	81.4
AD-SGAM	0.615	0.869	82.6	81.5

GP-SGAM:高斯过程平滑样条-空间广义加性模型 Gaussian process smooth spline-spatial generalized additive model; CR-SGAM:三次回归样条-空间广义加性模型 Cubic regression spline-spatial generalized additive model; TP-SGAM:薄板回归样条-空间广义加性模型 Thin plate regression spline-spatial generalized additive model; AD-SGAM:自适应样条-空间广义加性模型 Adaptive smoothing spline-spatial generalized additive model

表 5 Logistic 回归模型和 AD-SGAM 的几种指标拟合结果

Table 5 Fitting results of some indicators for Logistic regression model and AD-SGAM

模型检验指标 Model test index	逻辑斯蒂回归模型 Logistic regression model	空间广义加性模型 AD-SGAM
平均绝对误差 Mean absolute error	0.323	0.263
均方误差 Mean square error	0.162	0.132
均方根误差 Root mean squared error	0.402	0.363

### 2.4 黑龙江省林火发生概率

基于以上两种林火概率预测模型,通过预测火点与随机点的火灾发生概率,运用反距离权重插值法对黑龙江省林火空间概率分布进行插值分析,为增强林火火险区划精准性,减少非林业用地对结果的影响,剔除黑龙江省西南和东南地区主要农业用地。两种模型依据各自的阈值 0.475 和 0.615 以及 0.5 对黑龙江省进行火险等级划分,将林火发生概率小于阈值或 0.5 的区域设为低火险区,发生概率在阈值和 0.5 之间的区域设为中火险区,发生概率大于阈值或 0.5 的区域设为高火险区,最终生成了黑龙江省的森林火险区划图(图 5)。Logistic 回归模型的结果显示,黑龙江省除北部地区分布有中低火险区外几乎全省为林火高发地区,极少数林火中风险地区。而根据 AD-SGAM 的预测结果,黑龙江省中高火险区主要集中在西北地区以及中部和南部地区,北部地区分布着零星低火险区,东南部也有少量分布。

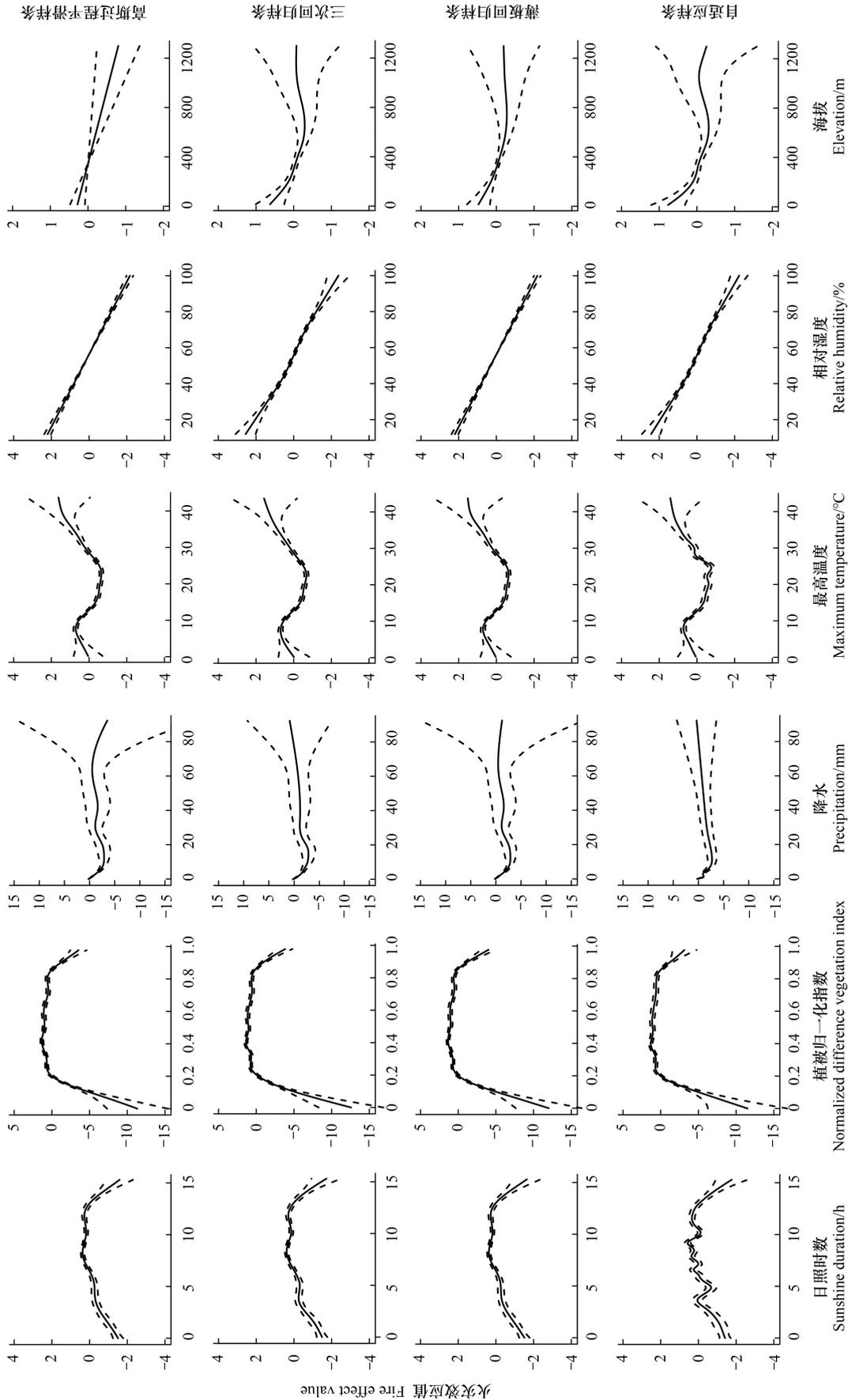


图3 使用四种平滑样条基拟合SGAM的一维非参数项部分依赖  
 Fig.3 One-dimensional partial dependence plots of SGAM fitted with four smoothing spline bases for non-parametric terms

虚线区域为95%置信区间。横轴表示自变量的值,纵轴表示 $\text{logit} = \{\log(P/(1-P))\}$ 的值,其中 $P$ 表示林火发生的概率

### 3 讨论

#### 3.1 黑龙江省森林火灾预测模型比较

本研究采用了常见的 Logistic 回归模型以及由 4 种平滑样条基(GP, CR, TP, AD)拟合的 SGAMs,对黑龙江省林火的驱动因子进行了对比研究。通过对每一个非参数项使用不同的平滑样条基进行组合,发现所有的非参数项使用同一平滑样条基拟合比使用不同的平滑样条基组合的拟合精度高,这与何培等<sup>[34]</sup>的研究结果一致。通过拟合结果的对比,AD 平滑样条基的拟合效果显著优于其他三种平滑样条基。Yang 等<sup>[35]</sup>通过使用自适应惩罚样条估计回归函数的研究同样得出了自适应惩罚样条优于传统光滑样条的结论。

Logistic 回归模型主要侧重于参数的估计与推断,建模时未充分考虑空间结构,并且忽略了林火发生过程中各解释变量之间存在的复杂关系,因此模型的拟合精度较低<sup>[15]</sup>。相比之下,SGAM 具有更强的灵活性和更优的拟合能力,能够通过引入非线性、交互项和分段函数,更好地描述林火发生过程中各驱动因子之间的非线性关系。同时,SGAM 将空间信息纳入模型,能够更准确地捕捉林火发生的空间变异性,从而有效提升预测精度<sup>[18]</sup>。

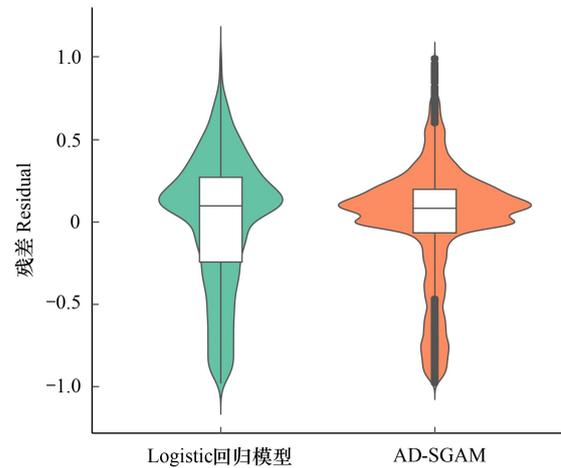


图 4 Logistic 回归模型与 AD-SGAM 的测试集残差分布

Fig. 4 Distribution of test set residuals for logistic regression model and AD-SGAM

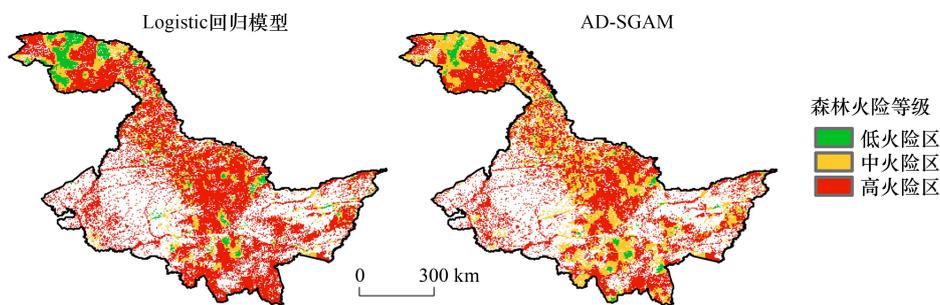


图 5 黑龙江省森林火险区划图

Fig. 5 Forest fire risk zoning map for heilongjiang province

#### 3.2 黑龙江省森林火灾发生的驱动因子分析

在解释 SGAM 结果时,重要的是综合考虑各个自变量的分段效应和影响强度。本研究基于 AD-SGAM 的拟合结果和部分依赖图,对黑龙江省的林火发生驱动因子的影响进行了分析。首先,“降水量”对森林火灾的影响呈现出先减弱后增强的趋势。这可能是因为适当的降水量会提高林间的相对湿度,促进地表植物的生长从而堆积了可燃物<sup>[36]</sup>。因此,相对湿度在低值时与林火发生呈正相关,而降水与林火发生的负相关关系逐渐减弱;当降水达到一定阈值,相对湿度超过约 55%时,有效抑制了林火的发生,以致相对湿度与林火发生呈负相关而降水与林火的负相关关系逐步递增。其次,NDVI 是反映植被生长的指标,当 NDVI 值小于 0.22 时,通常表示植被覆盖相对较少,可能是荒漠、裸露土壤或者衰老的植被<sup>[37]</sup>从而减少火灾发生的可能性。当 NDVI 值在 0.22—0.85 之间时,植被含水量增加,健康的中等丰度植被对森林火灾产生积极作用<sup>[38]</sup>。高丰度植被和燃料(NDVI>0.85)可能使森林保持较长时间的寒冷和高湿度环境<sup>[39]</sup>,从而减少火灾的发生。此外,低海拔地区往往聚集了大量人类活动,这增加了林火发生的风险。而随着海拔的增加,气候条件、植被覆盖和土壤湿度

逐渐不利于火灾的发生<sup>[40]</sup>。

“最高温度”、“日照时数”的影响趋势呈现复杂的曲线状,可能是因为这些气象因素与火灾发生之间存在复杂的非线性关系。

### 3.3 黑龙江省林火预测

从预测概率来看, Logistic 回归模型的预测概率范围为 0.022—0.947, 而 AD-SGAM 的预测概率范围为 0.043—0.936, 两者的差异较小。然而, 从预测准确性的角度来看, Logistic 回归模型在训练集和测试集上的预测准确率均低于 AD-SGAM。因此, Logistic 回归模型在黑龙江省森林火灾预测中的表现较差, 这与 Zhang 等<sup>[25]</sup>在其研究中的发现相符。根据全国森林防火规划, 黑龙江省的 1、2 级火险区主要集中在西北部、东部和南部<sup>[41]</sup>, 而 AD-SGAM 预测的中高火险区均被包含在其中。因此, SGAM 在拟合效果和预测准确度方面均优于传统 Logistic 回归模型, 这一结论与 Rodríguez-Pérez 等<sup>[18]</sup>的研究结果一致。

### 3.4 森林管理建议

本研究表明, 气象因素、可燃物因素和地形因素对森林火灾的影响呈现显著的分段效应。因此, 制定针对不同地区和时段的森林火灾管理措施至关重要。在植被稀少、低海拔的地区, 尤其是在相对湿度较低的时期, 应加强林业管理。具体措施包括加强现有树木的培育, 通过机械除草和定期清理枯枝落叶等方式减少可燃物的积累。同时, 通过植树造林和生态恢复等措施, 确保高密度的植被覆盖。此外, 定期进行人工灌溉可以提高稀疏林区的含水量, 增加森林湿度, 从而降低火灾风险。与此同时, 林业部门应在开展林业经济活动的过程中, 提高消防工作的效率和水平。在全球气候变化和经济快速发展的背景下, 森林经营战略需要更加注重气象、植被和地形等变量与森林火灾之间日益复杂的关系, 才能实现有效的森林火灾预防与控制。

### 3.5 研究局限性

本研究只考虑了空间经纬度之间的交互作用, 将 SGAM 中的自变量视为独立的影响因素, 未充分考虑自变量之间的相互作用效应, 这可能限制了模型在捕捉不同自变量之间关联效应方面的表现。在未来的研究中, 可以通过引入交互项, 更全面地捕捉各自变量间的复杂关系, 从而提升模型的准确性和表现力。此外, 森林火灾的发生不仅受自然因素影响, 还受到诸如人为活动等其他因素的影响。因此, 今后的研究应将自然因素与人为因素结合起来, 构建更加综合的模型, 以便对森林火灾的驱动机制进行更为详细的分析与研究。

## 4 结论

根据 2006—2020 年黑龙江省林火和气象数据、地形数据、植被覆盖数据, 本研究采用 Logistic 回归模型和 SGAM 对黑龙江省林火发生进行了预测, 并分析了各解释变量的影响。研究表明, 相较于 Logistic 回归模型, 采用四种不同平滑样条拟合的 SGAMs 均具有更高的预测精度。其中, AD 平滑样条基拟合的 SGAM 在拟合和预测上表现最优, 且能够提供对森林火灾事件的分段解释。因此, SGAM 在林火预测中具有重要的应用价值, 能够帮助森林管理者根据不同的时空条件优化预防措施。

### 参考文献 (References):

- [1] Venkatesh K, Preethi K, Ramesh H. Evaluating the effects of forest fire on water balance using fire susceptibility maps. *Ecological Indicators*, 2020, 110: 105856.
- [2] McKenzie D, Shankar U, Keane R E, Natasha Stavros E, Heilman W E, Fox D G, Riebau A C. Smoke consequences of new wildfire regimes driven by climate change. *Earth's Future*, 2014, 2(2): 35-59.
- [3] Graff C A, Coffield S R, Chen Y, Foufloula-Georgiou E, Randerson J T, Smyth P. Forecasting daily wildfire activity using Poisson regression. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(7): 4837-4851.
- [4] 孙龙, 尚喆超, 胡海清. Poisson 回归模型和负二项回归模型在林火预测领域的应用. *林业科学*, 2012, 48(5): 126-129.
- [5] Bugallo M, Esteban M D, Marey-Pérez M F, Morales D. Wildfire prediction using zero-inflated negative binomial mixed models: application to Spain. *Journal of Environmental Management*, 2023, 328: 116788.
- [6] 郭福涛, 苏漳文, 马祥庆, 宋禹辉, 孙龙, 胡海清, 杨婷婷. 大兴安岭塔河地区雷击火发生驱动因子综合分析. *生态学报*, 2015, 35(19): 6439-6448.
- [7] Sevinc V, Kucuk O, Goltas M. A Bayesian network model for prediction and analysis of possible forest fire causes. *Forest Ecology and Management*,

- 2020, 457: 117723.
- [ 8 ] Zema D A, Lucas-Borja M E, Fotia L, Rosaci D, Sarnè G M L, Zimbone S M. Predicting the hydrological response of a forest after wildfire and soil treatments using an Artificial Neural Network. *Computers and Electronics in Agriculture*, 2020, 170: 105280.
- [ 9 ] 周鹏飞, 王艳霞. 应用机器学习算法分析广西林火发生驱动因素及林火预测. *东北林业大学学报*, 2024, 52(11): 72-82.
- [ 10 ] 何锐, 陆恒, 晋子振, 秦艳, 杨弘毅, 刘治银, 杨广睿, 徐进滢, 龚雪, 赵求东. 基于随机森林算法的中国西南地区林火发生预测模型构建及驱动因子. *生态学报*, 2023, 43(22): 9356-9370.
- [ 11 ] 蔡奇均, 曾爱聪, 苏漳文, 郭福涛. 基于 Logistic 回归模型的浙江省林火发生驱动因子分析. *西北农林科技大学学报: 自然科学版*, 2020, 48(2): 102-109.
- [ 12 ] 张运林, 田玲玲, 丁波, 张延威, 刘讯, 武燕. 贵州省林火发生驱动因子及预测模型. *生态学杂志*, 2024, 43(1): 282-289.
- [ 13 ] Chang Y, Zhu Z L, Bu R C, Chen H W, Feng Y T, Li Y H, Hu Y M, Wang Z C. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landscape Ecology*, 2013, 28(10): 1989-2004.
- [ 14 ] Guo F T, Innes J L, Wang G Y, Ma X Q, Sun L, Hu H Q, Su Z W. Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *Journal of Plant Ecology*, 2015, 8(5): 480-490.
- [ 15 ] Guisan A, Edwards T C, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 2002, 157(2): 89-100.
- [ 16 ] 朱政, 赵璠, 王秋华, 邓小凡, 高仲亮. 林火发生预报模型研究进展. *世界林业研究*, 2022, 35(3): 26-31.
- [ 17 ] 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述. *计算机研究与发展*, 2019, 56(10): 2071-2096.
- [ 18 ] Rodríguez-Pérez J R, Ordóñez C, Roca-Pardiñas J, Vecín-Arias D, Castedo-Dorado F. Evaluating lightning-caused fire occurrence using spatial generalized additive models: a case study in central Spain. *Risk Analysis*, 2020, 40(7): 1418-1437.
- [ 19 ] Vilar L, Woolford D G, Martell D L, Martín M P. A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. *International Journal of Wildland Fire*, 2010, 19(3): 325.
- [ 20 ] Wood S N, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 2016, 111(516): 1548-1563.
- [ 21 ] Brillinger D R, Preisler H K, Benoit J W. Probabilistic risk assessment for wildfires. *Environmetrics*, 2006, 17(6): 623-633.
- [ 22 ] Parisien M A, Snetsinger S, Greenberg J A, Nelson C R, Schoennagel T, Dobrowski S Z, Moritz M A. Spatial variability in wildfire probability across the western United States. *International Journal of Wildland Fire*, 2012, 21(4): 313-327.
- [ 23 ] Woolford D G, Martell D L, McFayden C B, Evens J, Stacey A, Wotton B M, Boychuk D. The development and implementation of a human-caused wildland fire occurrence prediction system for the province of Ontario, Canada. *Canadian Journal of Forest Research*, 2021, 51(2): 303-325.
- [ 24 ] Phelps N, Woolford D G. Comparing calibrated statistical and machine learning methods for wildland fire occurrence prediction: a case study of human-caused fires in lac La biche, Alberta, Canada. *International Journal of Wildland Fire*, 2021, 30(11): 850-870.
- [ 25 ] Zhang Z, Yang S, Wang G Y, Wang W W, Xia H T, Sun S C, Guo F T. Evaluation of geographically weighted logistic model and mixed effect model in forest fire prediction in Northeast China. *Frontiers in Forests and Global Change*, 2022, 5: 1040408.
- [ 26 ] Justice C, Giglio L, Korontzi S, Owens J, Morisette J, Roy D, Descloitres J, Alleaume S, Petitcolin F, Kaufman Y. The MODIS fire products. *Remote Sensing of Environment*, 2002, 83(1): 244-262.
- [ 27 ] Edwards R B, Naylor R L, Higgins M M, Falcon W P. Causes of Indonesia's forest fires. *World Development*, 2020, 127: 104717.
- [ 28 ] Biswas S, Vadrevu K P, Lwin Z M, Lasko K, Justice C O. Factors controlling vegetation fires in protected and non-protected areas of Myanmar. *PLoS One*, 2015, 10(4): e0124346.
- [ 29 ] Syphard A D, Radeloff V C, Keuler N S, Taylor R S, Hawbaker T J, Stewart S I, Clayton M K. Predicting spatial patterns of fire on a southern California landscape. *International Journal of Wildland Fire*, 2008, 17(5): 602.
- [ 30 ] Yathish H, Athira K V, Preethi K, Pruthviraj U, Shetty A. A comparative analysis of forest fire risk zone mapping methods with expert knowledge. *Journal of the Indian Society of Remote Sensing*, 2019, 47(12): 2047-2060.
- [ 31 ] Pedersen E J, Miller D L, Simpson G L, Ross N. Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*, 2019, 7: e6876.
- [ 32 ] Wood S N, Augustin N H. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 2002, 157(2): 157-177.
- [ 33 ] Simpson G L. Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, 2018, 6: 149.
- [ 34 ] 何培, 辛士冬, 姜立春. 基于广义加性模型的樟子松树干削度方程研建. *北京林业大学学报*, 2020, 42(12): 1-8.
- [ 35 ] Yang L Q, Ding M Z, Hong Y M, Wang X J. Estimating functions and derivatives *via* adaptive penalized splines. *Communications in Statistics-Simulation and Computation*, 2021, 50(7): 1-18.
- [ 36 ] 张珍, 杨淞, 朱贺, 王光玉, 郭福涛, 孙帅超. 混合效应模型在林火发生预测中的适用性. *应用生态学报*, 2022, 33(6): 1547-1554.
- [ 37 ] Zhao L, Dai A G, Dong B. Changes in global vegetation activity and its driving factors during 1982—2013. *Agricultural and Forest Meteorology*, 2018, 249: 198-209.
- [ 38 ] Sow M, Mbow C, Hély C, Fensholt R, Sambou B. Estimation of herbaceous fuel moisture content using vegetation indices and land surface temperature from MODIS data. *Remote Sensing*, 2013, 5(6): 2617-2638.
- [ 39 ] Gabban A, San-Miguel-Ayanz J, Viegas D X. A comparative analysis of the use of NOAA-AVHRR NDVI and FWI data for forest fire risk assessment. *International Journal of Remote Sensing*, 2008, 29(19): 5677-5687.
- [ 40 ] Guo F T, Selvalakshmi S, Lin F F, Wang G Y, Wang W H, Su Z W, Liu A Q. Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Canadian Journal of Forest Research*, 2016, 46(4): 582-594.
- [ 41 ] 国家林业和草原局. 全国森林防火规划(2016—2025年). (2016-12-29) [2021-07-06]. <http://www.ndrc.gov.cn/fggz/fzslgh/gjjzqgh/201705/>